IES de Teis
Avda. de Galicia, 101
36216 – Vigo
886 12 04 64
ies.teis@edu.xunta.es

Busca en Internet información sobre estas cuestiones:

- La Inteligencia Artificial responsable es un tipo de IA que se encarga de pontenciar una serie de valores deseados en las sociedades, y que han sido recogidos por empresas como Microsoft y se hallan en la mayoría de tratados sobre la IA responsable, ¿cuáles serían esos valores o factores?

  Los valores y principios clave que suelen ser promovidos por la inteligencia artificial responsable incluyen:

  - **Transparencia:** La IA responsable busca ser transparente en sus procesos y decisiones, permitiendo a los usuarios comprender cómo se llega a ciertas conclusiones o recomendaciones.

  - **Equidad e inclusión:** Se esfuerza por garantizar que los sistemas de IA no perpetúen ni amplifiquen sesgos existentes en la sociedad, promoviendo la equidad y la inclusión de todas las personas, independientemente de su género, raza, etnia, orientación sexual, etc.

  - **Responsabilidad:** Los sistemas de IA deben ser diseñados para que las personas y organizaciones que los desarrollan sean responsables de sus impactos, ya sean positivos o negativos.

  - **Fiabilidad y seguridad:** La IA responsable se preocupa por garantizar que los sistemas sean confiables y seguros, minimizando los riesgos de errores o comportamientos inesperados que puedan causar daño.

  - **Privacidad y protección de datos:** Se enfoca en proteger la privacidad y los datos personales de los individuos, asegurando que la recopilación y el uso de datos se realicen de manera ética y conforme a las regulaciones pertinentes.

  - **Interpretación y explicación:** Busca proporcionar explicaciones claras y comprensibles sobre cómo funciona la IA y por qué toma ciertas decisiones, permitiendo a los usuarios evaluar su confiabilidad y tomar decisiones informadas.

  - **Colaboración humana:** Reconoce la importancia de la colaboración entre humanos y sistemas de IA, promoviendo la complementariedad entre las capacidades de ambos para lograr resultados óptimos.

- Los sesgos en la IA se pueden clasificar de distintas formas, pero resulta interesante el estudio desarrollado por dos investigadores del MIT, Harini Suresch y John Guttag, en el que identifican seis sesgos que deberían evitarse por los científicos de datos y expertos en IA, ¿cuáles son? Y descríbelos brevemente.

  Sesgo histórico

  Se considera que un modelo de inteligencia artificial produce sesgos históricos, cuando, aunque está entrenado con conjuntos de datos medibles y contrastados, produce resultados incorrectos, al interpretar el presente en términos que pueden pertenecer al pasado.

Este tipo de sesgo es el que provoca que algunos modelos de IA reproduzcan estereotipos que son dañinos para una parte de la población y que no se ajustan a la realidad. En el paper publicado por el MIT se explica por ejemplo cómo una institución financiera que utiliza un algoritmo IA para predecir la capacidad de pago de sus clientes, podría llegar a "aprender" que es mejor no conceder préstamos a personas afroamericanas, tomando como base únicamente los datos (y los estereotipos) con los que este algoritmo ha sido entrenado.

Sesgo de representación

El sesgo de representación se produce cuando los datos con lo que ha sido entrenado el modelo, infrarepresentan una parte de la población y como consecuencia, no es capaz de ofrecer una imagen amplia y diversa de la sociedad.

Un ejemplo de lo anterior es ImageNet, una enorme base de datos que contiene millones de imágenes etiquetadas y que habitualmente es utilizada para que los algoritmos de IA sean capaces de reconocer objetos. Pues bien, como apuntan desde el MIT, aproximadamente la mitad de estas imágenes se han tomado en Estados Unidos y en otros países occidentales y únicamente entre el 1% y el 2% de las imágenes representan culturas como la asiática, las distintas africanas y otras.

Como consecuencia, si por ejemplo preguntásemos a ImageNet por un vestido de novia, reconocería perfectamente el de un"novia occidental" (blanco), pero tendrá enormes dificultades para reconocer los trajes ceremoniales que se emplean en las bodas de Corea del Sur o Nigeria.

Sesgo de medición

El sesgo de medición se produce al elegir, recopilar o calcular las características y etiquetas que se utilizarán en un proceso de la predicción a futuro. Normalmente, se produce cuando se quiere analizar una característica o idea que no es directamente observable.

Por ejemplo, la «solvencia» de una persona es un concepto abstracto que a menudo se operativiza con un sustituto que puede ser medible, como la puntuación crediticia. En este caso, podemos encontrar un sesgo de medición cuando en el algoritmo se incluyen indicadores indirectos que no reflejan adecuadamente los distintos grupos que se analizan.

En el paper se pone como caso paradigmático de ese sesgo, el uso del polémico algoritmo COMPAS por el sistema judicial estadounidense. Supuestamente este algoritmo predice la probabilidad de que un acusado vuelva a delinquir y sus datos son utilizados por jueces y funcionarios de justicia para por ejemplo, decidir si un arrestado debe permanecer en prisión preventiva (y durante cuanto tiempo) o si más tarde, se le puede conceder la libertad condicional. Sin embargo y teniendo en cuenta que en Estados Unidos la población carcelaria en mayoritariamente afroamericana y latina, el algoritmo estima no tanto la posibilidad de un futuro delito, sino de un futuro arresto, introduciendo variables de mayor riesgo asociadas a la raza.

Sesgo de agregación

El sesgo de agregación surge cuando se utiliza un modelo único para datos en los que hay grupos subyacentes o tipos de datos que deberían considerarse de forma diferente. El sesgo de agregación se basa en la suposición de que la correspondencia entre las entradas y las etiquetas es coherente en todos los subconjuntos de datos cuando en realidad y a menudo, no es así.

Este tipo de sesgo conduce a un modelo que acabe no siendo óptimo para ninguno de los grupos que dice representar o, en muchos casos, que acabe representado únicamente al grupo dominante. A menudo este sesgo se produce cuando se toman los datos de las redes sociales como un todo (por ej, un hashtag de Twitter) sin tener en cuenta que son espacios en los que "conviven" culturas y espacios socio-demográficos muy diferentes.

Sesgo de Evaluación

Este sesgo suele desarrollarse cuando los datos de referencia utilizados para un objetivo concreto, no representa adecuadamente a la población objetivo que debe emplearse para el objetivo que se pretende.

Uno de los casos más paradigmáticos de sesgo de evaluación lo encontramos en distintos algoritmos de reconocimiento facial. Tal y como destacan los investigadores del MIT en este caso, algunos de estos algoritmos que claramente tienen un propósito comercial, no tienen ningún problema a la hora de identificar a varones blancos, pero presentan muchos más problemas cuando de lo que se trata es realizar un análisis sobre imágenes que representan a mujeres de piel oscura, confundiéndolas

con otras cosas. ¿Por qué? A menudo por la infrarepresentación de estas personas en los datos que se han utilizado para su entrenamiento.

Sesgo de despliegue

El sesgo de despliegue o implantación se produce cuando no hay correspondencia entre el problema que se pretende resolver con un modelo y la forma en que se utiliza en la práctica.

Esto ocurre a menudo cuando un sistema se construye y evalúa como si fuera totalmente autónomo, mientras que en realidad operan en un complicado sistema sociotécnico moderado por estructuras institucionales y responsables humanos. Así, los sistemas producen resultados que primero deben ser interpretados por los responsables humanos, para después llevar a la toma de decisiones. A pesar de su buen funcionamiento aislado, pueden acabar provocando consecuencias perjudiciales debido a fenómenos como la automatización o el sesgo de confirmación.

- La forma en la que la Unión Europea lucha contra el sesgo en la IA es a través de sus propias normativas que han permitido la creación de un marco de gobernanza para la inteligencia artificial, en el que los sistemas de IA se han dividido según cuatro niveles de riesgos, descríbelos brevemente.
  - **Riesgo bajo**: Se refiere a sistemas de IA que presentan un bajo riesgo para la seguridad y los derechos fundamentales de las personas. Ejemplos pueden incluir aplicaciones de IA para servicios de atención al cliente o sistemas de recomendación de productos. Estos sistemas están sujetos a una supervisión menos estricta y a menos requisitos regulatorios.
  - **Riesgo limitado**: Aquí se encuentran los sistemas de IA que presentan un riesgo limitado para la seguridad y los derechos fundamentales, pero aún así requieren un cierto nivel de regulación. Esto puede incluir sistemas de IA utilizados en el sector de la salud para ayudar en el diagnóstico médico, donde la precisión y la seguridad son de suma importancia.
  - **Riesgo alto**: Se refiere a sistemas de IA que presentan un alto riesgo para la seguridad y los derechos fundamentales de las personas. Estos sistemas pueden tener un impacto significativo en la vida de las personas o en la sociedad en general, como los sistemas de IA utilizados en la gestión de recursos humanos que podrían perpetuar sesgos o discriminación. Están sujetos a regulaciones más estrictas y a un escrutinio más detallado.
  - **Riesgo inaceptable**: Son los sistemas de IA que se consideran una amenaza para la seguridad y los derechos fundamentales de las personas y, por lo tanto, su uso puede ser prohibido. Esto incluiría sistemas de IA que manipulan el comportamiento humano de manera inaceptable o que violan la privacidad de las personas de manera significativa.

- Un ejemplo de IA explicable sería en el campo de la medicina, cuando vamos a consulta médica o a urgencias, ¿sería bueno una IA capaz de explicarnos todo lo que hay tras unas decisiones o predicciones? Razona tu respuesta.

Considero positivo que pueda haber mas opciones de consultas para responder dudas o para hacer predicciones, etc… Pero siempre en compañía de la opinión, explicación por parte de un médico, es decir que no se delegue en IAs la totalidad de las decisiones/explicaciones si no que debería utilizarse como una herramienta mas a la que acudir por parte de los médicos.

- En un análisis sobre hacia dónde va la IA, la empresa Gartner ha manifestado que la innovación en la IA va a un ritmo rápido y que debemos seguir a ese paso acelerado, a través de cuatro tendencias que irán marcando la agenda y la ejecución de las empresas, ¿Cuáles son esas cuatro tendencias? Descríbelas brevemente.

IES de Teis
Avda. de Galicia, 101
36216 – Vigo
886 12 04 64
ies.teis@edu.xunta.es

- **Automatización inteligente**: Esta tendencia se refiere a la integración de la inteligencia artificial en los procesos de automatización empresarial para mejorar la eficiencia, la precisión y la escalabilidad. La automatización inteligente utiliza tecnologías como el aprendizaje automático y la automatización robótica de procesos (RPA) para optimizar tareas y procesos comerciales.

- **IA responsable**: Con el aumento del uso de la inteligencia artificial en diversas aplicaciones, surge la necesidad de abordar preocupaciones éticas y sociales relacionadas con su desarrollo y despliegue. La IA responsable se centra en garantizar que los sistemas de IA sean éticos, justos, transparentes y confiables, protegiendo los derechos y el bienestar de las personas.

- **IA aumentada**: Esta tendencia implica la combinación de capacidades humanas y de IA para mejorar la productividad y la toma de decisiones. La IA aumentada incluye tecnologías como la realidad aumentada, la realidad virtual y los asistentes virtuales inteligentes que ayudan a los humanos a realizar tareas de manera más eficiente y efectiva.

- **IA democratizada**: La democratización de la inteligencia artificial busca hacer que la IA sea más accesible y utilizada por una amplia gama de personas y organizaciones. Esto implica simplificar el desarrollo, la implementación y el uso de la inteligencia artificial mediante herramientas y plataformas accesibles, así como la capacitación y educación en IA para una mayor adopción y comprensión.

-